

The Flow of Biotechnology Information

Gene

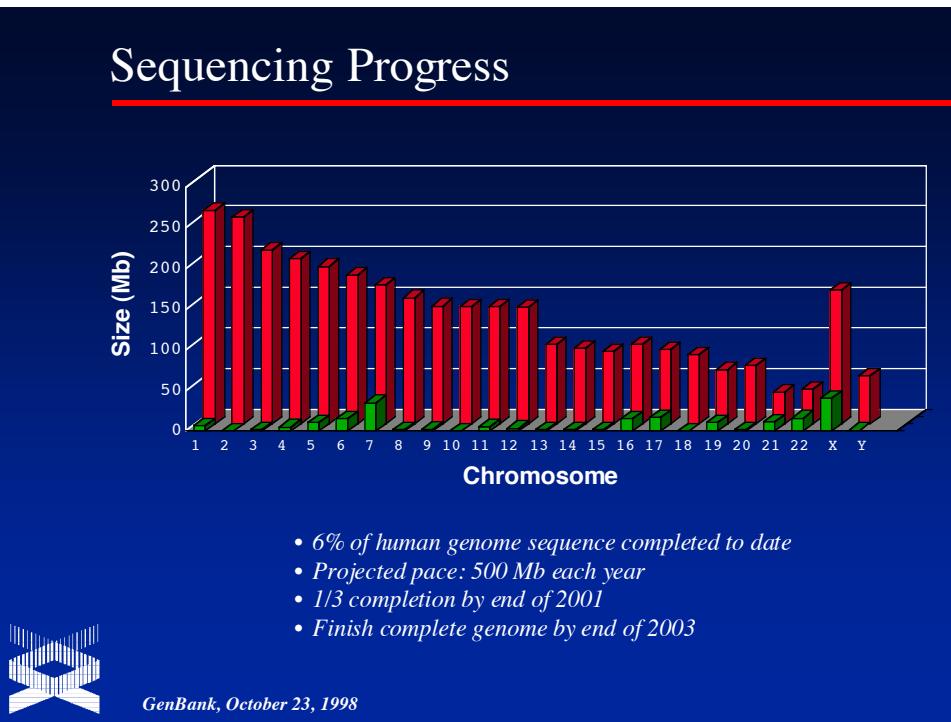


Function



> DNA sequence
AATTCCATGAAATCGTATACTGGTCTGGTACCGGAAACAC
TGAGAAAATGGCAGAGCTCATCGCTAAAGGTATCATCGAA
TCCTGGTAAAGACGTCAACACCATCAACGTCTTGACGTTA
ACATCGATGAACACTCTGAACGAAGATAACCTGATCTCGG
TTGCTCTGCCATGGCGGATGAAGTTCTGAGGAAGCGAA
TTTGAAACCGTTCATCGAAGAGATCTCTACCAAATCTCG
GTAAGAAAGTTGCCCTCTTCGTTCTTACGGTTGGGGGA
CGGTAAGTGGATGCGTGACTTCTGAAGAACGTATGAAACGGC
TACGGTTGGTTGTGTGTGAGACCCGCTGATGTTTCAGA
ACGAGGCCGACCGAAGCTGAGCAGGACTGCTATCGAATTGG
TAAGAAGATCGCGAACATCTAGTAGA



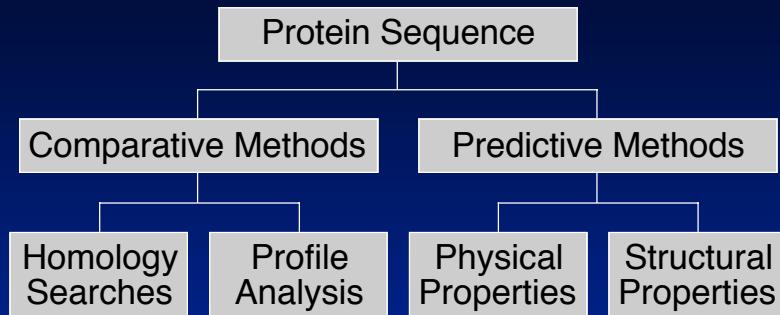


Protein Conformation

- Christian Anfinsen
Studies on reversible denaturation →
“Sequence specifies conformation”
- Chaperones and disulfide
interchange enzymes:
involved but not controlling final state
- “Starting with a newly-determined sequence, what
can be determined computationally about its
possible function and structure?”



Protein Sequence Analysis



- *Shared ancestry?*
- *Similar function?*
- *Domain or complete sequence?*



BLAST

- Seeks high-scoring segment pairs (HSP)
 - pair of sequences that can be aligned without gaps
 - when aligned, have maximal aggregate score
(score cannot be improved by extension or trimming)
 - score must be above score threshold S
 - gapped (2.0) or ungapped (1.4)
- Search engines
 - WWW search form
<http://www.ncbi.nlm.nih.gov/BLAST>
 - Unix command line
`blastall -p programe -d db -i query > outfile`
 - E-mail server
`blast@ncbi.nlm.nih.gov`



BLAST Algorithms

<i>Program</i>	<i>Query Sequence</i>	<i>Target Sequence</i>
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Nucleotide, six-frame translation	Protein
TBLASTN	Protein	Nucleotide, six-frame translation
TBLASTX	Nucleotide, six-frame translation	Nucleotide, six-frame translation



Neighborhood Words

Query Word (W = 3)

Query: GSQSLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNIVEAFVED

Neighborhood
Words

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc.	

Neighborhood Score
Threshold
(T = 13)



High-Scoring Segment Pairs

PQG	18
PEG	15
PRG	14
PKG	14
PNG	13
PDG	13
PHG	13
PMG	13
PSG	13
PQA	12
PQN	12
etc.	

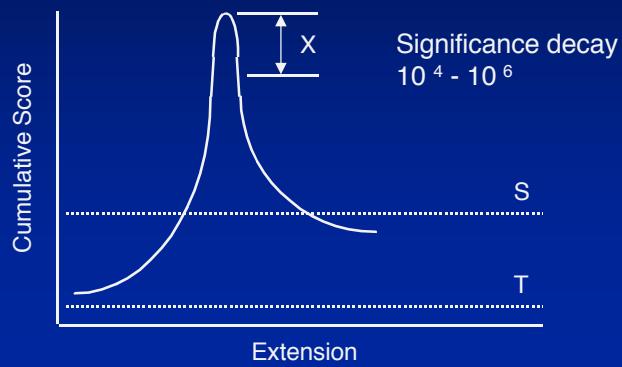


Query: 325 SLAALLNKCKTPQGQRLVNQWIKQPLMDKNRIEERLNIVEA 365
+LA++L TP G R++ +W+ P+ D + ER + A
Sbjct: 290 TLASVLDCTVTPMGSRMLKRWLHMPVRDTRVLLERQQTIGA 330



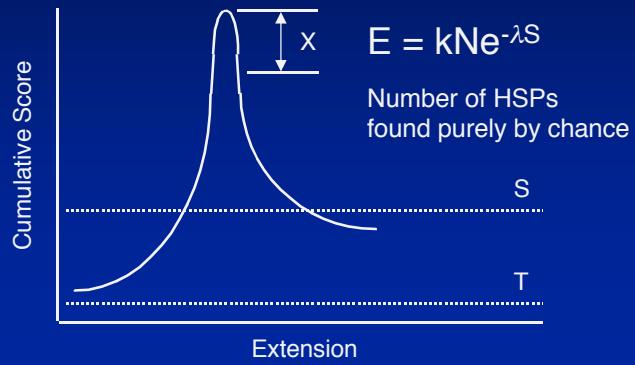
BLAST Search Requirements

- A query sequence, in FASTA format
- Which BLAST program to use
- Which database to search
- Parameter values



BLAST Search Requirements

- A query sequence, in FASTA format
- Which BLAST program to use
- Which database to search
- Parameter values



Scoring Matrices

- Empirical weighting scheme to represent biology
 - Cys/Pro important for structure and function
 - Trp has bulky side chain
 - Lys/Arg have positively-charged side chains
- Importance of understanding scoring matrices
 - Appear in all analyses involving sequence comparison
 - Implicitly represent a particular theory of evolution
 - Choice of matrix can strongly influence outcomes



Matrix Structure

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	-2	2	-3	0	0	-1
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-2	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4			
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4		
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1



PAM Matrices

- Margaret Dayhoff, 1978
- Point Accepted Mutation (PAM)
 - Look at patterns of substitutions in related proteins
 - The new side chain must function the same way as the old one (“acceptance”)
 - On average, 1 PAM corresponds to 1 amino acid change per 100 residues
 - 1 PAM ~ 1% divergence
 - Extrapolate to predict patterns at longer distances



PAM Matrices

- Assumptions
 - Replacement is independent of surrounding residues
 - Sequences being compared are of average composition
 - All sites are equally mutable
- Sources of error
 - Small, globular proteins used to derive matrices (departure from average composition)
 - Errors in PAM 1 are magnified up to PAM 250
 - Does not account for conserved blocks or motifs



BLOSUM Matrices

- Henikoff and Henikoff, 1992
- **Blocks Substitution Matrix (BLOSUM)**
 - Look only for differences in conserved, ungapped regions of a protein family
 - More sensitive to structural or functional substitutions
 - BLOSUM n
 - Contribution of sequences $>n\%$ identical weighted to 1
 - Reduces contribution of closely-related sequences
 - Increasing $n \sim$ increasing PAM distance



So many matrices...

- Triple-PAM strategy (*Altschul, 1991*)
 - PAM 40 Short alignments, highly similar
 - PAM 120 ↓
 - PAM 250 Longer, weaker local alignments
- BLOSUM 62 (*Henikoff, 1993*)
 - Most effective in detecting known members of a protein family
 - BLAST default
- No single matrix is the complete answer for all sequence comparisons



BLAST Query

```
>N-terminal unknown protein
MSSAAAAAGAGGGALFQPOSVSTANSNNNNNTPAALATHSPTNSNPVSGASSASSLLTAAGGNL
FGSSAKMLNELFGRQMKQAQDATSLPQSLDNAMLAAAMETATSAELIGSLNSTSKLLQQQHNNN...
```

↓ *BLASTP / SWISSPROT / BLOSUM62*

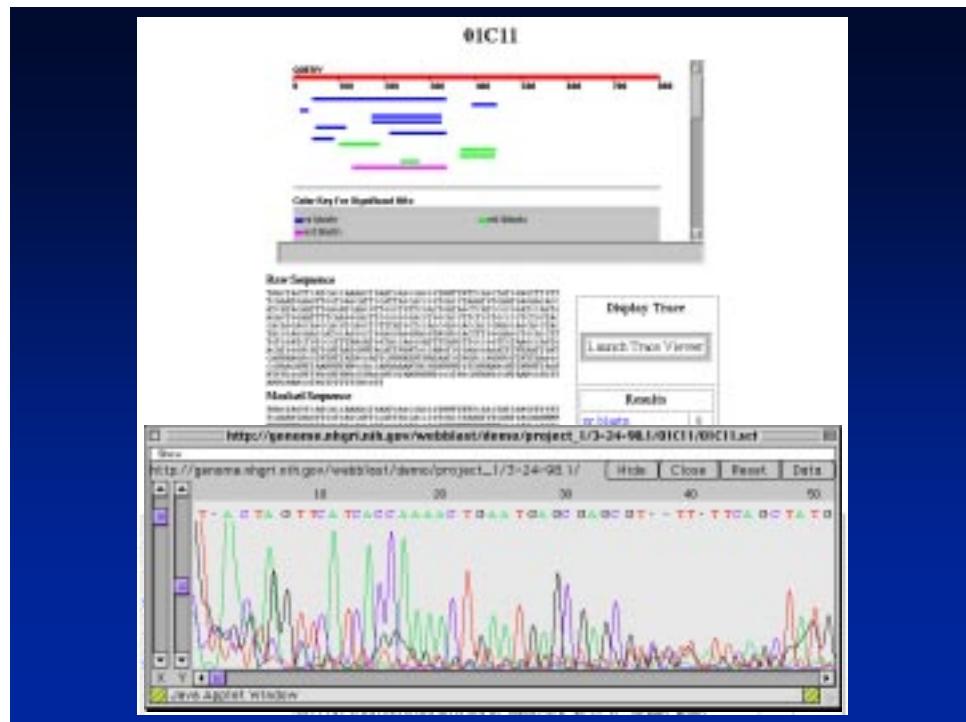
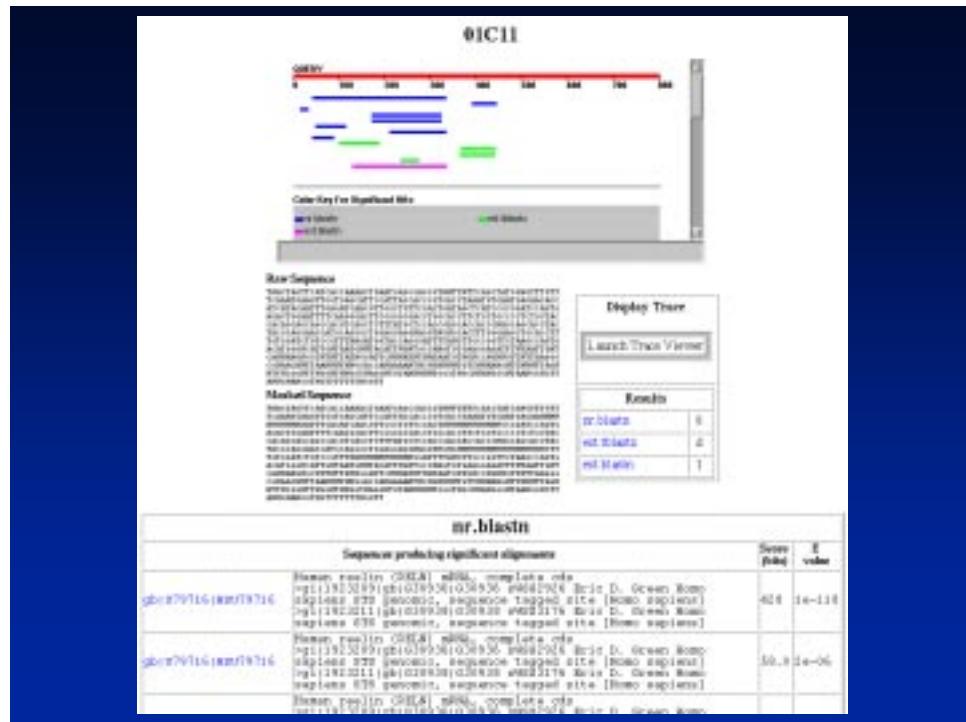
	Score (bits)	E Value
Sequences producing significant alignments:		
sp P29617 PRO_DROME PROTEIN PROSPERO	948	0.0
sp P34522 HM26 CAEEL HOMEOBOX PROTEIN CEH-26	242	4e-63
sp P48437 PRX1_MOUSE HOMEOBOX PROSPERO-LIKE PROTEIN PROX1 (PROX 1)	214	7e-55
sp Q92786 PRX1_HUMAN HOMEOBOX PROSPERO-LIKE PROTEIN PROX1 (PROX 1)	214	7e-55
sp Q91018 PRX1_CHICK HOMEOBOX PROSPERO-LIKE PROTEIN PROX1 (PROX 1)	213	2e-54
sp P25440 RNG3_HUMAN RING3 PROTEIN (KIAA9001)	35	0.79
sp P31000 VIME_RAT VIMENTIN	34	1.4
sp P48670 VIME_CRIGR VIMENTIN	34	1.4

Lower probability infers greater significance – but always look at the alignments!

WebBLAST

- Impetus
 - Need to archive data in a logical fashion
 - Shortcomings of commercial LIMS products
 - Need to perform many BLAST searches (locally)
- Goals
 - Collect and organize sequence data
 - Provide automated BLAST runs
 - Monthly re-BLAST against NCBI-month
 - Combine data from multiple sources
 - Allow for export to assembly programs
 - Use in multi-user, multi-project environment
 - Most steps transparent to users

Current Topics in Genome Analysis '99
Protein Sequence Analysis: BLAST and Beyond



Profiles

- Numerical representations of multiple sequence alignments
- Depend upon *patterns* or *motifs* containing conserved residues
- Represent the common characteristics of a protein family
- Can find similarities between sequences with little or no sequence identity
- Allow for the analysis of distantly-related proteins



Profile Construction

```

APHIIIVATPG
GCEIVIAATPG
GVEICIAATPG
GVDILIGATPG
RPHIIIVAATPG
KPHIIIAATPG
KVQLIILATPG
RPDIVIAATPG
APHIIIVGATPG
APHIIIVGATPG
GCHVVIAATPG
NQDIVVVAATPG

```

- Which residues are seen at each position?
- What is the frequency of observed residues?
- Which positions are conserved?
- Where can gaps be introduced?

Position-Specific Scoring Table

Cons	A	B	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y	Z
G	17	18	0	19	14	-22	31	0	-9	12	-15	-5	15	10	9	6	18	14	1	-15	-22	11
P	18	0	13	0	0	-12	13	0	8	-3	-3	-1	-2	23	2	-2	12	11	17	-31	-8	1
H	5	24	-12	29	25	-20	8	32	-9	9	-10	-9	22	7	30	10	0	4	-8	-20	-7	27
I	-1	-12	6	-13	-11	33	-12	-13	63	-11	40	29	-15	-9	-14	-15	-6	7	50	-17	8	-11
V	3	-11	1	-11	-9	22	-3	-11	46	-9	37	30	-13	-3	-9	-13	-6	6	50	-19	2	-8
V	5	-9	9	-9	-9	19	-1	-13	57	-9	35	26	-13	-2	-11	-13	-4	9	58	-29	0	-9
A	54	15	12	20	17	-24	44	-6	-4	-1	-11	-5	12	19	9	-13	21	19	9	-39	-20	10
T	40	20	20	20	20	-30	40	-10	20	20	-10	0	20	30	-10	-10	30	150	20	-60	-30	10
P	31	6	7	6	6	-41	19	11	-9	6	-16	-11	0	89	17	17	24	22	9	-50	-48	12
G	70	60	20	70	50	-60	150	-20	-30	-10	-50	-30	40	30	20	-30	60	40	20	-100	-70	30



ProfileScan

- Search sequence against a collection of profiles
- Databases available
 - PROSITE 1167 entries
 - Pfam 527 entries
- [http://www.ch.embnet.org/software/
PFSCAN_form.html](http://www.ch.embnet.org/software/PFSCAN_form.html)



ProfileScan Query

```
>c-terminal end
MALLQISEPGLSAAPHQRRRLAAGIDLGTTNSLVTVRSGQAETLADHEGRHLLPSVVHYQQQGHGSVGYDA
RTNAALDTANTISVVKRLMGRSLADIQQRYPHLPPQFQASENGLPMIETAAGLLNPVRVSADILKALAAR
ATEALAGELDGVVITVPAYFDDAQQRQGTKDAARLAGLHVRLRNNEPTAAATAYGLDSQEGVIAVYDLGG
GTFDISILRLSRGVFEVLATGGDSALGGDFDHLLADYIREQAGIPDRSDNRVQRELLDAAIAAKIA...
```

↓ Prosite + Pfam
↓ Significant matches only

normalized raw	from	- to	Profile	Description
219.3535	21	600	PF00012	HSP70 Heat shock hsp70 proteins

↓ E-value

NScore	SwissProt
7.0	1.8000
8.0	0.1800
9.0	0.0180
10.0	0.0018
219.4	3e-211

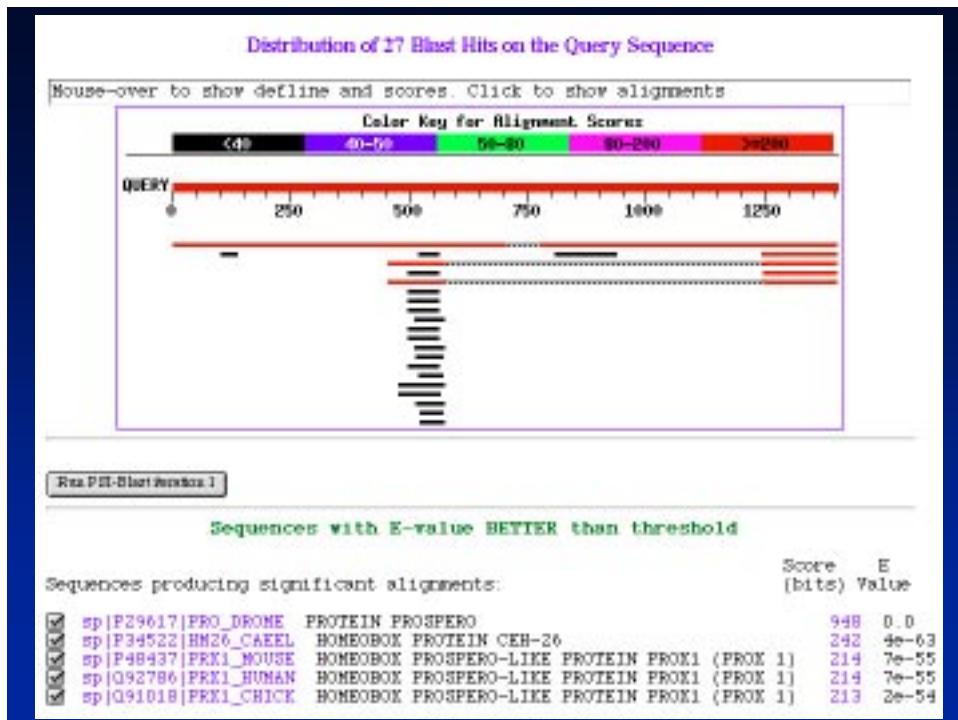
↓ Signatures

[IV]-D-L-G-T-[ST]-x-[SC]
[LIVMF]-[LIVMFY]-[DN]-[LIVMFS]-G-[GSH]-[GS]-[AST]-x(3)-[ST]-[LIVM]-[LIVMFC]
[LIVM]-x-[LIVMF]-x-G-G-x-[ST]-x-[LIVM]-P-x-[LIVM]-x-[DEQKRSTA]



PSI-BLAST

- Position-Specific Iterated BLAST search
- Easy-to-use version of a profile-based search
 - Perform BLAST search against protein database
 - Use results to calculate a position-specific scoring matrix
 - PSSM replaces query for next round of searches
 - May be iterated until no new significant alignments are found
 - Convergence – all related sequences deemed found
 - Divergence – query is too broad, make cutoffs more stringent



BLOCKS

- Steve Henikoff, Fred Hutchinson Cancer Research Center, Seattle
- Multiple alignments of conserved regions in protein families
 - 1 “block” = 1 short, **ungapped** multiple alignment
 - Families can be defined by one or more blocks
 - Searches allow detection of one or more blocks representing a family
- Search engines
 - E-Mail *blocks@howard.fhcrc.org*
 - Web *http://blocks.fhcrc.org/*



BLOCKS Query

```
>C-terminal end
MALLQISEPGLSAAPHQRRRLAAGIDLGTTNSLVATVRSQQAETLADHEGRHLLPSVVHYQQQGHSGVYDA
RTNAALDTANTISSVKRLLMGRSLADIQQRYPHLPPQFQASENGLPMIETAAGLLNPVRVSADILKALAAR
ATEALAGELDGVVITVPAYFDDAQRQGTKDAARLAGLHVLRLLNEPTAAATAYGLDSQEGVIAVYDLGG
GTFDISILRLSRGVFEVLATGGSALGGDFDHLLADYIREQAGIPDRSDNRVQRELLDAAIAAKIA...
```

↓ *Search blocks*

BL00297A HSCA_ECOLI 136	ALAARATEALAGELDGVVITVPAYFDDAQRQGTKDAARLAGLHVLRLLNEPTAAA
C-terminal 136	ALAARATEALAGELDGVVITVPAYFDDAQRQGTKDAARLAGLHVLRLLNEPTAAA

↓ *Examine blocks* □□

```
ID  HSP70_1; BLOCK
AC  BL00297A; distance from previous block=(94,187)
DE  Heat shock hsp70 proteins family proteins.
BL  PRR motif; width=55; seqs=111; 99.5%=2947; strength=1607
```



BLOCKS Entry

```

ID  HSP70_1; BLOCK
AC  BL00297A; distance from previous block=(94,187)
DE  Heat shock hsp70 proteins family proteins.
BL  PRR motif; width=55; seqs=111; 99.5% = 2947; strength=1607
HS70_CHLRE ( 129) KETAQASLGADREVKKAVVTVPAYFNDSQRQATKDAGMIAGLEVLRIINEPTAAA 19
HS7L_SBYV ( 132) ALISTASEAFKCQCTGVICSVNPANYNCLQRSFTESCVNLSGYPCVYMVNEPSAAA 75
HS7R_HUMAN ( 124) KLKETAESVILKPPVVDCVVSVPFCYTDAERRSVM DATQIAGLNCLRLMNETTAVA 45
HS7T_MOUSE ( 126) TKMKETAEVFWAPMSQRVITVPAYFNDSQRQATKDAGVIAGLNLVRLIINEPTAVA 28
YKH3_YEAST ( 160) SLLKDRDARTEDFVNKMSFTIPDFFDQHQRKALLDASSITTGIEETYLVSEGMSV 100
DNAK_BACSU ( 95) HLKSYAESYLGETVSKAVITVPAYFNDQERQATKDAGKIAGLEVERIINEPTAAA 7
DNAK_BORBU ( 122) KMKETAEVAYLGEKVTCAVITVPAYFNDQERQATKDAGKIAGLEVKRIVNEPTAAA 3
DNAK_BRUOV ( 122) KMKETAESYLYGETVTQAVITVPAYFNDQERQATKDAGKIAGLEVKRIIINEPTAAA 3
DNAK_BURCE ( 123) KMKTAEDYLVLEPVTECAVITVPAYFNDQERQATKDAGRIAGLEVVKRIIINEPTAAA 3
DNAK_CAUCR ( 122) KMKEAAAHLGEPVTKAVIDVPAYFNDQERQATKDAGKIAGLEVLRIIINEPTAAA 5
DNAK_CHLPN ( 125) KMKETAEVAYLGETVTEAVITVPAYFNDQERQASTKDAGRIAGLDVKRIIPEPTAAA 10
DNAK_CLOPE ( 98) KLLKADAEAYLGEKVTCAVITVPAYFNDQERQATKDAGRIAGLDVKTIINEPTAAS 8
DNAK_CRYPH ( 122) KLVVDASKYLGESVKQAVITVPAYFNDQERQATKDAGRIAGLEVVKRIIINEPTAAS 5
DNAK_ECOLI ( 121) KMKTTAEDYLVLEPVTECAVITVPAYFNDQERQATKDAGRIAGLEVVKRIIINEPTAAA 3
DNAK_ERYRH ( 96) YMKSYAEDYLVLEKVTKAVIDVPAYFNDQERQATKDAGKIAGLEVERIINEPTAAA 5
DNAK_HAEIN ( 120) KMKTTAEDFLGESVTECAVITVPAYFNDQERQATIDAGKIAGLDVVKRIIINEPTAAA 6
.
.
.

```



BLOCK Maker

```

>chk-H5
SRRSASHPTYSSEMIAAIARAEKSRGSSRQSIQKYIKSHYKVGHNADLQIKLSIRRLLAAGVLIKQTKGVGASGSFRIAKS
>hum-H1
TPRKASGPVSELITKAVAAASKERSGVSLAALKKALAAAGYDVKEKNNNSRICKLGLKSLVSKGTLVQTKGTGASGSFKLNKK
>pea-H1
PRNPASHPTYEIMKDAIVSLKEKNQSSQYAIAKFIEEKQKQLPANFKKLQLQNLKKNVASGKLIVKGSFKLSAAAKKP

```

↓ MOTIF/GIBBS

```

>Histone chk-H5 family
6 sequences are included in 2 blocks

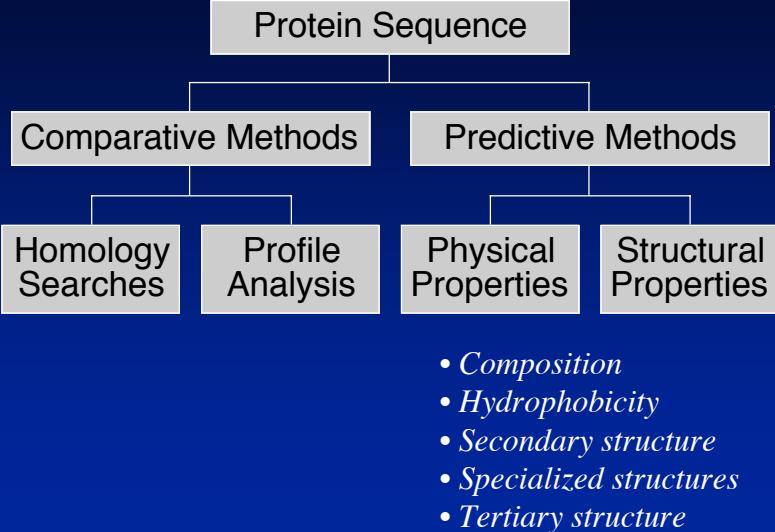
HistoneA, width = 31
chk-H5      1 SHPTYSEMIAAAIRAEKSRGSSRQSIQKYI
hum-H1      1 SGPPVSELITKAVAAASKERSGVSLAALKKAL
pea-H1      1 SHPTYREMIMKDAIVSLKEKNQSSQYAIAKFIE
sce-H1.1    1 SSKSYSRELIIEGLTALKERKGSSRPALKKFI
sce-H1.2    1 SSLTYKEMILKSMPOQLNDKGSSSRIVLKKYV
xla-H1      1 SGPSASELIVKAVSSSKERSGVSLAALKKAL

HistoneB, width = 15
chk-H5 ( 21) 53 IRRLLAAGVLIKQTKG
hum-H1 ( 21) 53 LKSLVSKGTLVQTKG
pea-H1 ( 21) 53 LKKNVASGKLIVKVG
sce-H1.1 ( 21) 53 IKKGEAGDFEQPKG
sce-H1.2 ( 21) 53 IKKCVENELVQPKG
xla-H1 ( 21) 53 LKALVTKGTLTQVKG

```



Protein Sequence Analysis



ProtParam

- Computes physicochemical parameters
 - Molecular weight
 - Theoretical pI
 - Amino acid composition
 - Extinction coefficient
- Simple query
 - SWISS-PROT accession number
 - User-entered sequence, in single-letter format
- <http://expasy.hcuge.ch/sprot/protparam.html>



ProtParam Query

```
MNGEADCPTDLEMAAPKGQDRWSQEDMLTLLECMKNLPSNDSSKFKTTESHMDWEKVAFKDFSGDMCKLK  
WVEISNEVRKFRTLTTELILDAQEHVKNPYKGKKLKKHPDPKKPLTPYFRFFMEKRAKYAKLHPEM...
```

↓ *Compute parameters*

Number of amino acids: 727			
Molecular weight: 84936.8			
Theoretical pI: 5.44			
Amino acid composition:			
Ala (A) 35	4.8%	Leu (L) 57	7.8%
Arg (R) 39	5.4%	Lys (K) 97	13.3%
Asn (N) 28	3.9%	Met (M) 25	3.4%
Asp (D) 58	8.0%	Phe (F) 18	2.5%
Cys (C) 6	0.8%	Pro (P) 39	5.4%
Glu (Q) 36	5.0%	Ser (S) 67	9.2%
Glu (E) 98	13.5%	Thr (T) 22	3.0%
Gly (G) 26	3.6%	Trp (W) 11	1.5%
His (H) 11	1.5%	Tyr (Y) 20	2.8%
Ile (I) 18	2.5%	Val (V) 16	2.2%
Asx (B) 0	0.0%		
Glx (Z) 0	0.0%		
Xaa (X) 0	0.0%		
Total number of negatively charged residues (Asp + Glu): 156			
Total number of positively charged residues (Arg + Lys): 136			



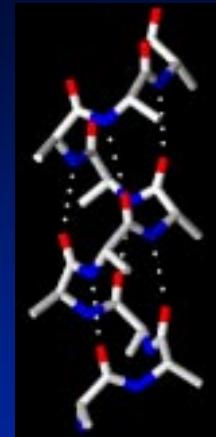
Secondary Structure Prediction

- Deduce the most likely position of alpha-helices and beta-strands
- Confirm structural or functional relationships when sequence similarity is weak
- Determine guidelines for rational selection of specific mutants for further laboratory study
- Basis for further structure-based studies



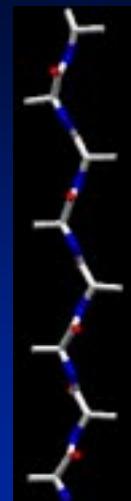
Alpha-helix

- Corkscrew
- Main chain forms backbone, side chains project out
- Hydrogen bonds between CO group at n and NH group at $n+4$
- Helix-formers: Ala, Glu, Leu, Met
- Helix-breaker: Pro



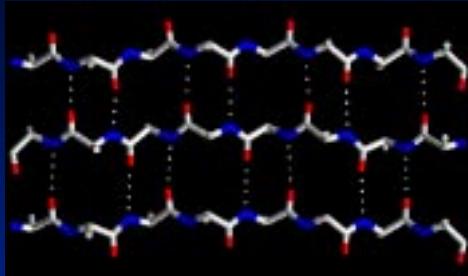
Beta-strand

- Extended structure (“pleated”)
- Peptide bonds point in opposite directions
- Side chains point in opposite directions
- No hydrogen bonding *within* strand



Beta-sheet

- Stabilization through hydrogen bonding
- Parallel or antiparallel
- Variant: beta-turn



Folding Classes



α

Cyt c

β

CD4

$\alpha+\beta$

Staph
nuclease

α/β

Triose
phosphate
isomerase

Globins

Orthogonal

EF-hand

Up-Down

Cytochrome

Orthogonal

Super-barrel

Greek key

Sandwich

Jelly roll

Split sandwich

Meander

Metal-rich

Open roll

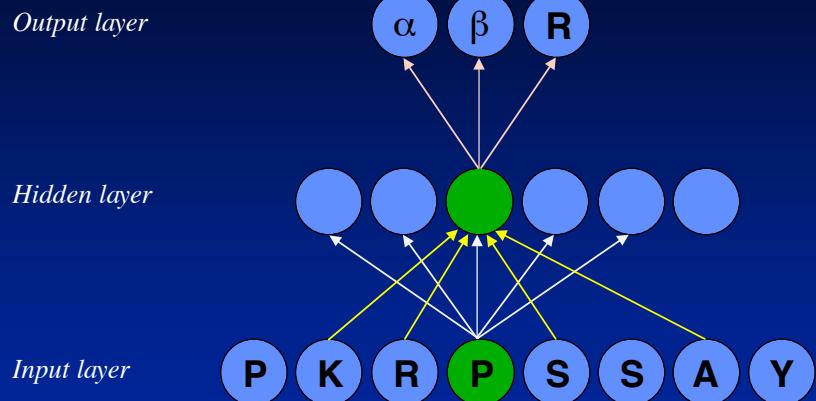
OB/UB roll

TIM barrel

Doubly-wound



Neural Network



PredictProtein

- Multi-step predictive algorithm (*Rost et al., 1994*)
 - Protein sequence queried against SWISS-PROT
 - MaxHom used to generate iterative, profile-based multiple sequence alignment (*Sander and Schneider, 1991*)
 - Multiple alignment fed into neural network (PHDsec)
- Accuracy
 - Average $> 70\%$
 - Best-case $> 90\%$
- Search engines
 - E-mail predictprotein@embl-heidelberg.de
 - Web <http://www.embl-heidelberg.de/predictprotein/>

PredictProtein Query

Joe Buzzcut
National Human Genome Research Institute, NIH
buzzcut@nhgri.nih.gov
flavodoxin - *Anacystis nidulans*
AKIGLFYGTQTCVQTQTCIQQEFGGSESIVDLNDIANADASDLNAYDYLIIGCPTWNVGEQLQSDWEGIY
BDDLDLVSNFQGKVAxFGAGDSQVGSDFQDAMGILEEKISSLGSQTVGYWPIEGYDFNESKAVRNQNQFVG
LAIDEDNOPLDTKKNR1KTWSVOLSQKSFGL

- SWISS-PROT hits
 - Multiple alignment
 - PDB homologues



Accuracy of Predictions



SignalP

- Neural network trained based on phylogeny
 - Gram-negative prokaryotic
 - Gram-positive prokaryotic
 - Eukaryotic
- Predicts secretory signal peptides
(*not* those involved in intracellular signal transduction)
- <http://www.cbs.dtu.dk/services/SignalP>



SignalP Query

```
>sp|P05019|IGFB_HUMAN INSULIN-LIKE GROWTH FACTOR IB PRECURSOR
MGKISSLPTQLFKCCFCDFLKVKMHTMSSHLFYLALCLLTFTSSATAGPETLCGAEVDAQFVCGDRG
```

↓
N-terminal end only
Eukaryotic set

```
***** SignalP predictions *****
Using networks trained on euk data

>IGF-IB      length = 195

# pos  aa    C      S      Y
.
.
.
46   A    0.365  0.823  0.495
47   T    0.450  0.654  0.577
48   A    0.176  0.564  0.369
49   G    0.925  0.205  0.855
50   P    0.185  0.163  0.376
```

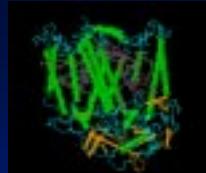
C = cleavage site score
S = signal peptide score
Y = combined score

```
.
```

```
< Is the sequence a signal peptide?
# Measure Position Value Cutoff Conclusion
max. C 49 0.925 0.37 YES
max. Y 49 0.855 0.34 YES
max. S 37 0.973 0.88 YES
mean S 1-48 0.550 0.48 YES
# Most likely cleavage site between pos. 48 and 49: ATA-GP
```



Transmembrane Classes



- Helix bundles
Long stretches of apolar amino acids
Fold into transmembrane alpha-helices
“Positive-inside rule”

*Cell surface receptors
Ion channels
Active and passive transporters*



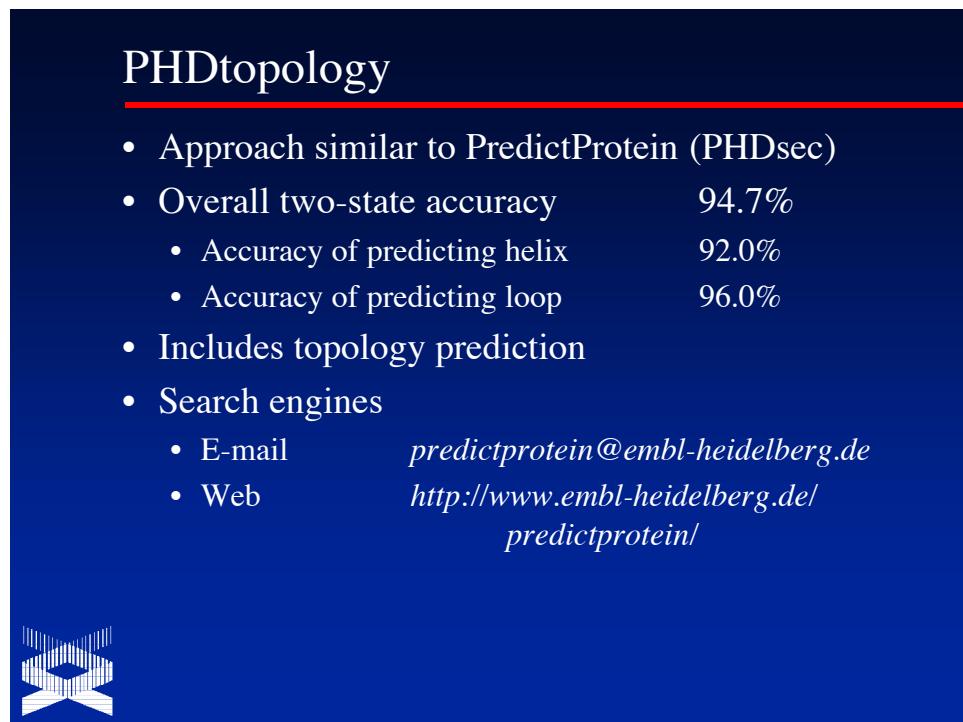
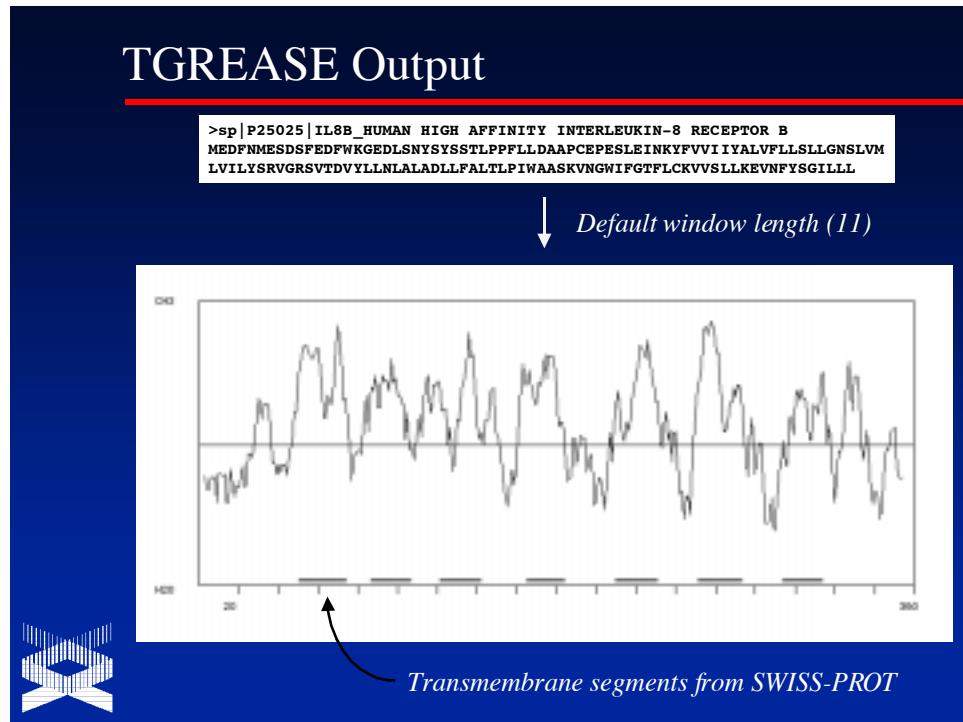
- Beta-barrel
Anti-parallel sheets rolled into cylinder
*Outer membrane of Gram-negative bacteria
Porins (passive, selective diffusion)*



TGREASE

- Calculates hydrophobicity along length of a protein
(Kyte and Doolittle, 1982)
- Hydropathy scale
 - Propensity to bury side chain within protein core
 - Based on solubility, free energy of transfer through water-vapor transition, and other factors
 - More positive scores indicate greater hydrophobicity
 - More negative scores indicate greater hydrophilicity
- Moving average, with 7-11 residues optimal
- <ftp://ftp.virginia.edu/pub/fasta>





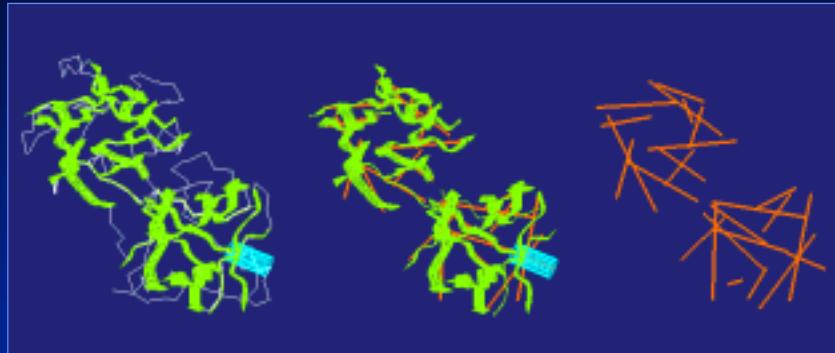
PHDtopology Query

Predicting Tertiary Structure

- Sequence specifies conformation, *but* conformation does *not* specify sequence
 - Structure is conserved to a much greater extent than sequence
 - Limited number of protein folds
 - Similarities between proteins may not necessarily be detected through “traditional” methods
 - Protein folding problem
 - Asilomar structure prediction “contest”
 - Numerous protein folds can be reliably identified
 - Consensus approach

VAST Structure Comparison

Step 1: Construct vectors for secondary structure elements

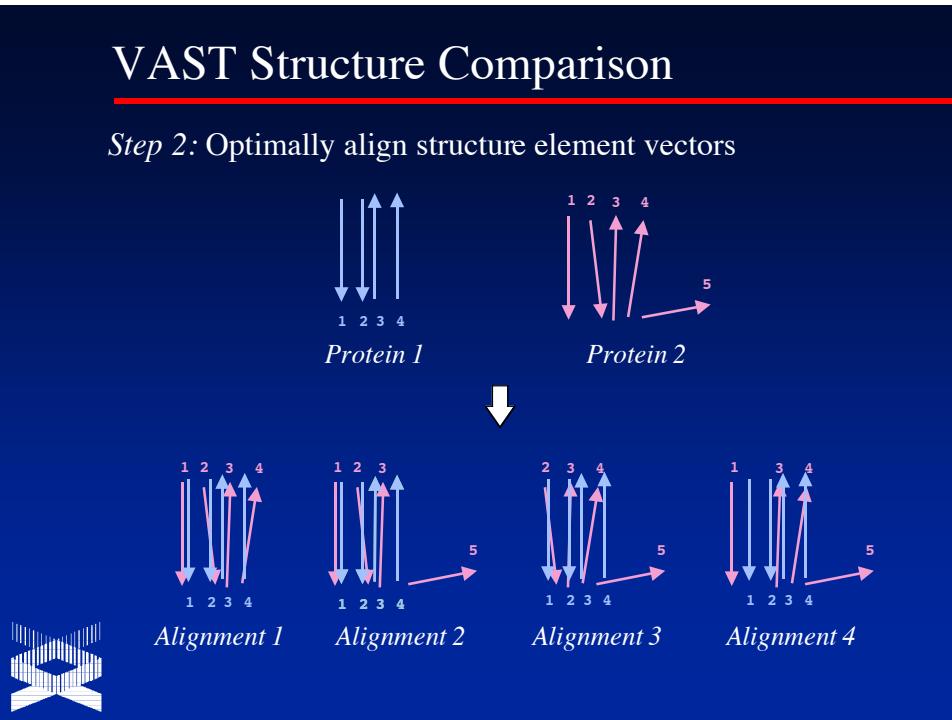


Ricin Chain B



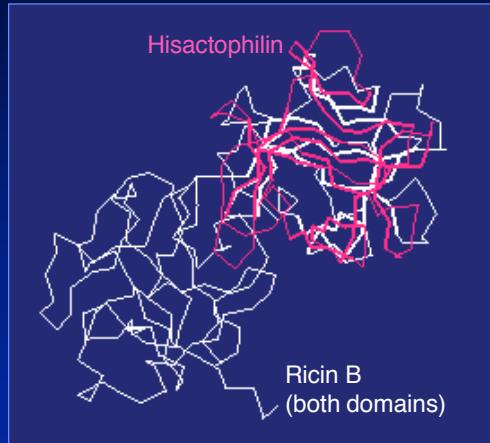
VAST Structure Comparison

Step 2: Optimally align structure element vectors



VAST Structure Comparison

Step 3: Refine residue-by-residue alignment using Monte Carlo



Hisactophilin

Ricin B
(both domains)

MMDB Id: 2778 PDB Id: 2LIV

Protein Chains: (single chain)
MEDLINE: PubMed
Taxonomy: Escherichia coli

PDB Authors: J.S Sack, M.A Saper & F.A Quiocho
PDB Deposition: 10-Apr-99
PDB Class: Periplasmic Binding Protein
PDB Compound: Leucine[Slash]Isoleucine[Slash]Valine-Binding Protein (LIVBP)

Sequence Neighbors: (single chain)
Structure Neighbors: (single chain), 1, 2

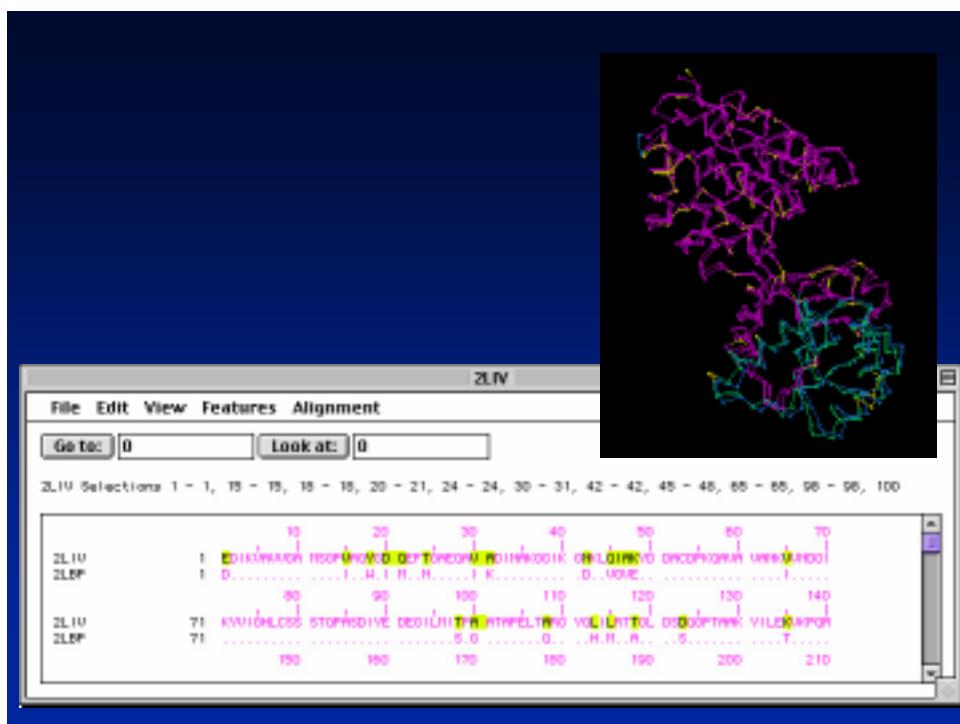
[View / Save Structure](#) [Get Cn3D 2.0 Now!](#)

Options: Viewer: Complexity:

Launch Viewer Cn3D v2.0 (zen.1) Cn3D Subset Up to 5 Models
 See File Cn3D v1.0 (zen.1) Virtual Bond Model Up to 10 Models
 Save File Mage All Atom Model All Models
 RasMol (PDB)

Current Topics in Genome Analysis '99 *Protein Sequence Analysis: BLAST and Beyond*

	PDB	C	D	SCo	P-VAL	RMSD	NRES	#id	Description
■	ZLBP	1		22.8	10e-18.7	0.9	269	76.4	Leucine-Binding Protein (LBP)
■	IPEA	1		20.7	10e-14.3	3.1	217	14.3	Amide Receptor NEGATIVE REGULATOR OF THE AMIDASE OPERON OF Pseudomonas aeruginosa (Amic) Complexed With Acetamide
■	IBNT_A_Z		2	15.1	10e-7.4	2.9	120	8.3	Methionine Synthase (B12-Binding Domains) (EC.2.1.1.19)
■	IDHR			17.1	10e-7.0	4.3	111	10.8	Dihydropteridine Reductase (DHPr) (EC.1.6.99.10) Complex With NADH
■	BABP	1		14.9	10e-7.0	3.2	125	10.4	L-Arabinose-Binding Protein (Mutant With Met 108 Replaced By Leu) (M108L) Complex With D-Galactose
■	ISCV_A_Z		2	14.5	10e-7.0	2.5	101	10.9	Succinyl-CoA Synthetase (Succinate-CoA Ligase) (Adp-Forming) (EC.6.2.1.5)
■	ZLBP	2		14.4	10e-6.7	2.8	110	10.0	Leucine-Binding Protein (LBP)
■	SCUT			13.5	10e-6.1	3.0	114	7.9	Cutinase (EC.3.1.1.-) Complexed With The Inhibitor Diethyl Para-Nitrophenyl Phosphate



SWISS-MODEL

- Automated comparative protein modelling server
- <http://www.expasy.ch/swissmod/SWISS-MODEL.html>

Results returned by E-mail

BLAST search to find similarities in PDB *by sequence*

↓

Select templates with sequence identity > 25% and projected model size > 20 amino acids

↓

Generate models

↓

Do energy minimization

↓

Generate PDB file for new protein model

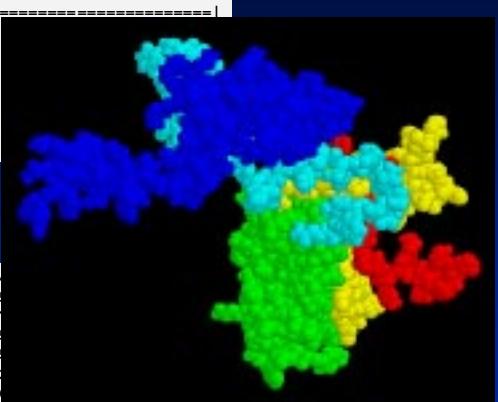


```
21DJH.pdb: 42.77 % identity
21DJG.pdb: 42.77 % identity
11DJG.pdb: 42.22 % identity
11QAS.pdb: 44.17 % identity
11QAT.pdb: 43.52 % identity
21QAT.pdb: 43.52 % identity
21QAS.pdb: 43.52 % identity
```

Target:

21DJH.pdb	-----
21DJG.pdb	-----
11DJG.pdb	-----
11QAS.pdb	-----
11QAT.pdb	-----
21QAT.pdb	-----
21QAS.pdb	-----



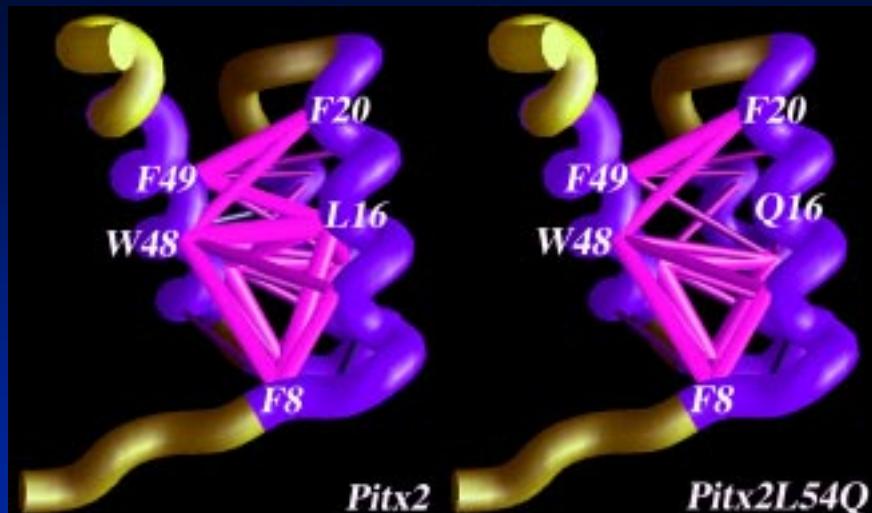
ATOM	1	H1	SER	1	24.219	22.95
ATOM	2	H2	SER	1	24.770	21.43
ATOM	3	N	SER	1	24.355	22.18
ATOM	4	H3	SER	1	23.466	21.92
ATOM	5	CA	SER	1	25.266	22.67
ATOM	6	CB	SER	1	24.826	24.07
ATOM	7	OG	SER	1	24.857	25.00
ATOM	8	HG	SER	1	24.717	25.929 -55.233 1.00 99.00
ATOM	9	C	SER	1	25.471	21.750 -53.751 1.00 25.00
ATOM	10	O	SER	1	25.923	22.169 -52.684 1.00 25.00
ATOM	11	N	LYS	2	25.227	20.460 -53.972 1.00 25.00
ATOM	12	H	LYS	2	24.961	20.142 -54.878 1.00 99.00
ATOM	13	CA	LYS	2	25.366	19.408 -52.943 1.00 25.00
ATOM	14	CB	LYS	2	24.003	18.772 -52.622 1.00 25.00

Rieger Syndrome / Iridogoniodysgenesis

- Map locus 4q25-26
- Gene encodes a bicoid-class homeodomain protein
- RS and IGD caused by mutations in same gene
- Slate-grey or chocolate brown eye color
- RS: White line on posterior cornea (Schwalbe's line): dense collagen and spindle-shaped cells with a basement membrane
- IGD: Stroma of iris hypoplastic and light in color
- Progression to glaucoma



Threading Analysis



$\Delta\Delta G = 8 \text{ kcal/mol}$ (20% difference)

